УДК 621.396.9

СРАВНЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ РАДИОЛОКАЦИОННЫХ ИЗОБРАЖЕНИЙ ОБЪЕКТОВ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТЬЮ И СЕТЬЮ С АРХИТЕКТУРОЙ VISION TRANSFORMER

Купряшкин И.Ф., *д.т.н., доцент, начальник кафедры «Военного учебно-научного центра Военно-воздушных сил «Военно-воздушная академия им. проф. Н.Е. Жуковского и Ю.А. Гагарина» (г. Воронеж), e-mail: ifk78@mail.ru*

CONVOLUTIONAL NEURAL NETWORK AND A NETWORK WITH VISION TRANSFORMER ARCHITECTURE ACCURACY COMPARISON IN RADAR IMAGES CLASSIFICATION TASK

Kupryashkin I.F.

Training and test data preparing procedure is described, a structures of a deep convolutional neural network and a network with the Vision Transformer are given. Their accuracy in classifying MSTAR dataset radar images at different noise levels was compared. It is shown that the use of the Vision Transformer architecture in this task does not lead to an increase in classification accuracy even with a significant change in the hyperparameters.

Key words: deep convolutional neural network, transformer, radar image, classification accuracy.

Ключевые слова: глубокая сверточная нейронная сеть, трансформер, радиолокационное изображение, точность классификации.

Описан порядок подготовки обучающих и тестовых данных, приведены структуры глубокой сверточной нейронной сети и сети с архитектурой Vision Transformer. Проведена их сравнительная оценка по точности классификации отметок объектов на радиолокационных изображениях набора MSTAR при различном уровне шума. Показано, что применение архитектуры Vision Transformer в рассматриваемой задаче не приводит к увеличению точности классификации даже при существенном изменении гиперпараметров.

Введение

Вопросы обработки радиолокационных изображений (РЛИ) земной поверхности и объектов с использованием технологий глубокого обучения продолжают вызывать значительный интерес и во многом уже достаточно хорошо проработаны [1-8]. Как правило, в числе наиболее перспективных инструментов для решения задачи классификации РЛИ объектов до недавнего времени рассматривались глубокие сверточные нейронные сети (ГСНС).

Однако начиная с 2017 года, с момента публикации работы [9] активное развитие получили сети, основанные на архитектуре Transformer, использующей механизмы внимания (self-attention). На начальном этапе новая архитектура продемонстрировала наиболее впечатляющие результаты в задачах обработки естественного языка, и достаточно отметить, что все современные большие языковые модели основаны именно на ней. Эти успехи вполне ожидаемо стимулировали применение архитектуры Transformer в других задачах, в том числе задачах компьютерного зрения.

Одно из первых описаний эффективного варианта трансформера для обработки изображений приведено в [10]. Предложенная архитектура получила название Vision Transformer (ViT) и в настоящий момент сеть OmniVec [11], основанная на ViT, демонстрирует наилучшую точность (92,4 % Top-1 Accuracy) на наборе ImageNet. К настоящему времени уже известны различные варианты реализации трансформеров для обработки РЛИ [12-14], однако исследование их возможностей при различном уровне шума на изображении не проводилось.

В связи с этим сравнение возможностей ViT-архитектуры и ГСНС в задачах обработки радиолокационных изображений представляет безусловный интерес, и целью настоящей работы является сравнение точности классификации малоразмерных объектов на детальных радиолокационных изображениях, достигаемой при использовании указанных архитектур, при различном уровне шума.

Характеристика обучающих и тестовых наборов данных

В [15] подробно описан порядок подготовки обучающих и тестовых наборов данных, и в настоящей работе он является полностью аналогичным. Так же в качестве исходных данных использована открытая часть набора MSTAR [16], включающего радиолокационные изображения X-диапазона десяти образцов военной техники с разрешением 0,3 м (рисунок 1).

Весь имеющийся набор изображений разбит на обучающий (1923 изображения, угол визирования при съемке 17 градусов), проверочный (891 изображение, угол визирования 17 градусов) и тестовый (2503 изображения, угол визирования 15 градусов) наборы.

Каждое изображение набора MSTAR хранится в отдельном файле в форме двумерного массива комплексных отсчетов $\dot{\mathbf{I}}$. Предварительная обработка этих изображений включает подавление фона местности вблизи отметок объектов, понижение размерности изображений до 44×44 отсчетов, добавление аддитивного гауссовского шума для имитации воздействия шумовой

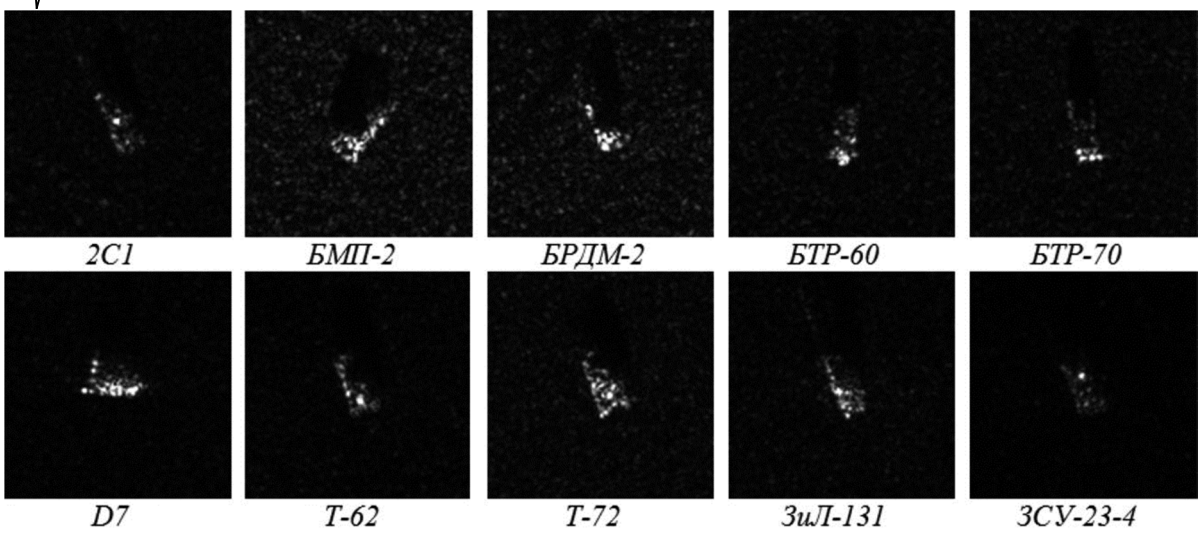


Рисунок 1 – Радиолокационные изображения объектов набора MSTAR

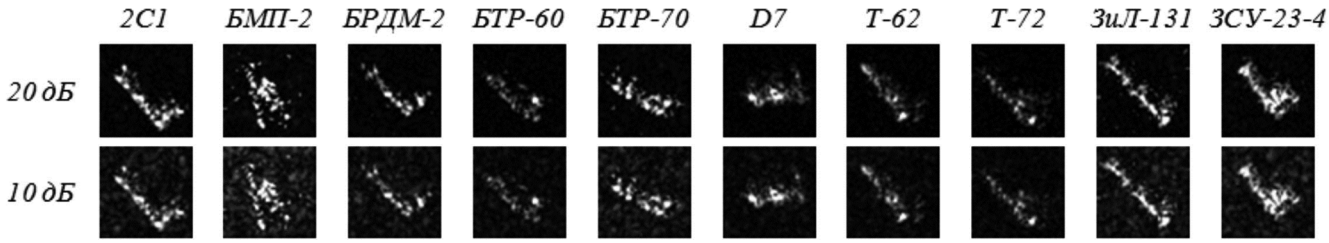


Рисунок 2 – Примеры изображений обучающих наборов

помехи на РЛС, пространственную фильтрацию и преобразование в целочисленный беззнаковый восьмибитный формат [15, 17].

Сформированы два одинаковых набора изображений, отличающихся только средним уровнем шумовой помехи. В первом наборе уровень шума выбирался таким, чтобы отношение среднего уровня отметок объектов к среднему уровню помехи на изображении составило $q = 20$ дБ, во втором наборе – $q = 10$ дБ. Примеры изображений приведены на рисунке 2.

Архитектура нейронных сетей и результаты их обучения

В качестве инструмента для разработки и обучения сетей использовался фреймворк TensorFlow 2.0 и библиотека Keras [18].

Для сверточной сети использовалась архитектура VGG-типа [19], включающая четыре пары сверточных слоев с одинаковыми ядрами 3×3, единичным шагом свертки и количеством фильтров 32, 64, 128 и 256 в каждой паре соответственно. После каждой пары включены слои подвыборок 2×2. Классификатор включает полносвязный входной слой на 4096 входов и 256 выходов, слой 50-процентного прореживания и выходной полносвязный слой с 10 выходами по числу классов объектов. В качестве функции активации сверточных слоев и входного слоя классификатора используется ReLU. Размерность входного слоя составляла 128×128 отсчетов. Количество настраиваемых параметров сети составило 2'223'082. Архитектура сверточной сети представлена на рисунок 3,*а*.

Архитектура трансформера включает входной слой с размерностью 128×128, к которому подключен слой, обеспечивающий преобразование входного изображения в последовательность его фрагментов (patches). Длина последовательности определяется исходя из заданной размерности фрагмента, так, при заданной в рассматриваемом случае размерности отдельного фрагмента 32×32 отсчета их количество составляет 16.

Далее каждый фрагмент преобразуется в одномерный вектор длиной 1024 элемента, и подается на позиционный кодировщик (patch encoder), включающий полносвязный слой и слой кодирования позиции (эмбеддинга). Входом полносвязного слоя является собственно вектор, входом слоя эмбеддинга – номер соответствующего фрагмента в поступающей последовательности. Сумма их выходов является общим выходом позиционного кодировщика. Размерность выхода является гиперпараметром сети, в данном случае ее значение задавалось равным 128.

Закодированная последовательность фрагментов изображения поступает на стек из восьми одинаковых блоков трансформера. Каждый блок трансформера включает слой нормализации, слой многоголового внимания (с размерностью векторов запроса, ключа и значения, равной 96, и четырьмя «головами» внимания), полносвязные слои с 10-и процентным прореживанием. Количество полносвязных слоев составляет 16 и определяется длиной последовательности, в каждом из них осуществляется обработка отдельных кодированных векторов последовательности независимо друг от друга. В качестве функции активации полносвязных слоев применяется geLU [20].

Выходом стека блоков трансформера является последовательность из 16 векторов с длиной 128 элементов каждый, которая перед подачей на вход полно-

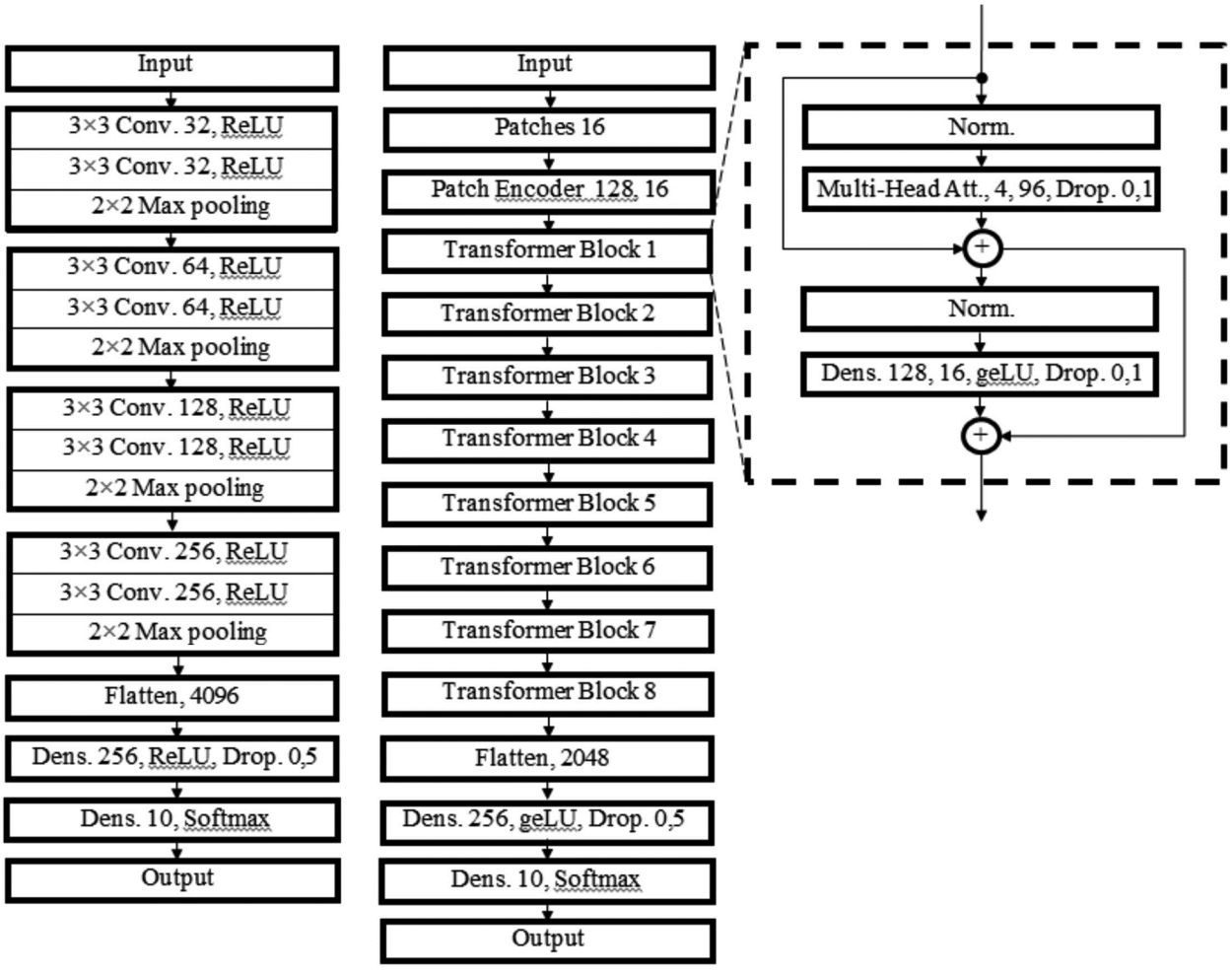
связного классификатора преобразуется в единый вектор длиной 2048. Классификатор включает входной полносвязный слой на 2048 входов и 256 выходов с функцией активации geLU, слой 50-процентного прореживания и выходной полносвязный слой с 10 выходами по числу классов объектов с функцией активации softmax.

Гиперпараметры трансформера выбирались таким образом, чтобы общее количество настраиваемых параметров (обучаемых весов), равное 2'379'914, примерно соответствовало количеству настраиваемых весов сверточной нейросети (2'223'082) для обеспечения возможности объективного сравнения получаемых результатов. Архитектура трансформера представлена на рисунке 3,*б*.

Условия обучения сверточной сети и трансформера являются одинаковыми. В качестве оптимизатора используется адаптивный алгоритм Adam с параметром скорости обучения 10^{-4}, обучение осуществляется в течение четырехсот (для наборов с $q = 20$ дБ) и шестисот (для наборов с $q = 10$ дБ) эпох пакетами по 64 изображения. В качестве функции потерь применена перекрестная энтропия [21]. Общее время обучения с использованием GPU (GeForce GTX1650) составляет примерно 27 минут ($q = 20$ дБ, 400 эпох) и 40 минут ($q = 10$ дБ, 600 эпох) как для сверточной сети, так и для трансформера, т.е. около четырех секунд на эпоху во всех случаях. В качестве дополнительного приема предотвращения переобучения применяется расширение данных (augmentation) путем смещения каждого изображения обучающего набора по вертикали и горизонтали на случайное число пикселей от одного до пяти. Этот же прием при подаче изображений валидационного и тестового наборов не применялся. Перед подачей на вход нейронной сети целочисленные значения диапазона [0; 255] преобразуются в вещественные диапазона [0; 1] путем деления на 255.

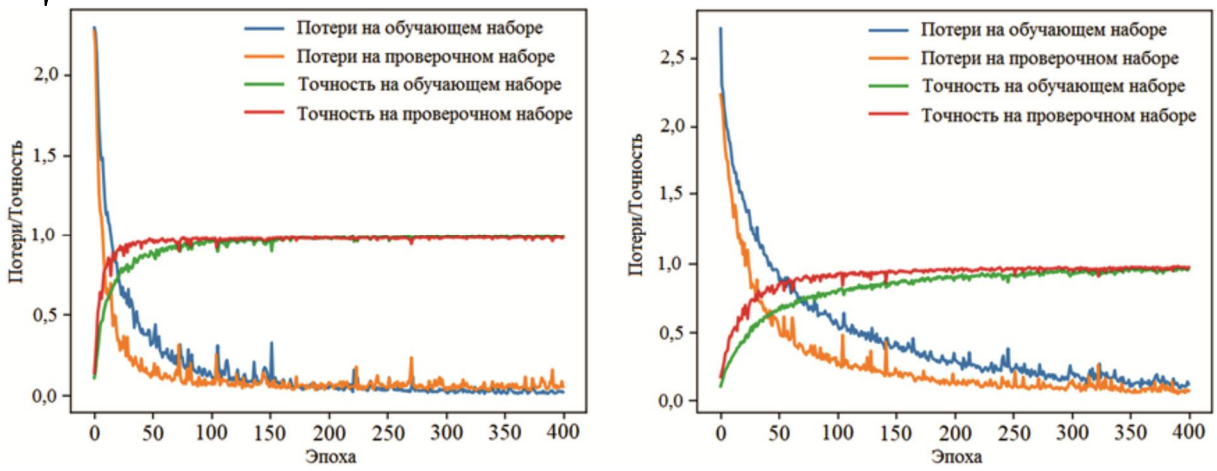
Ход процесса обучения на наборе с $q = 20$ дБ иллюстрируют зависимости функции потерь и точности классификации на рисунке 4, на наборе с $q = 10$ дБ – аналогичные зависимости на рисунке 5.

На тестовом наборе с $q = 20$ дБ сверточная сеть продемонстрировала точность 97,84 %, а трансформер – 95,43 %. Полученная оценка точности трансформера хорошо согласуется с результатами 95,96 % и 95,61 %, полученными в [13] и [14] соответственно при сходных, хотя и не полностью аналогичных, условиях обучения. На наборе с $q = 10$ дБ результат составил 95,59 % и 92,79 % соответственно. То есть трансформер с сопоставимым количеством весов на обоих наборах продемонстрировал несколько худшие результаты. Кроме того, из анализа зависимостей на рисунках 4 и 5 следует, что сверточная сеть характеризуется еще и более



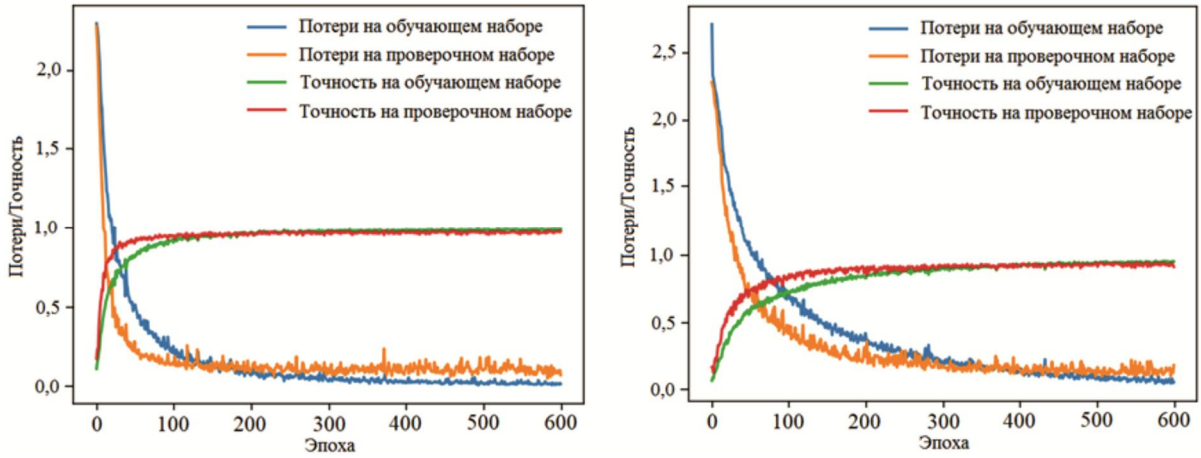
а) *б)*

Рисунок 3 – Архитектура сверточной нейросети (а) и трансформера (б)



а) *б)*

Рисунок 4 – Ход процесса обучения сверточной (а) сети и трансформера (б) при q = 20 дБ



а) *б)*

Рисунок 5 – Ход процесса обучения сверточной (а) сети и трансформера (б) при q = 10 дБ

высокой скоростью обучения (скоростью изменения точности и потерь). Можно также отметить несколько лучшую стабильность процесса обучения обоих сетей на наборе с $q = 10$ дБ, что обусловлено, по всей видимости, регуляризирующим эффектом шумовой помехи. Особенность зависимостей, состоящая в более быстром нарастании точности не на обучающем, а на валидационном (проверочном) наборе, обусловлена отсутствием случайных сдвигов изображений последнего при подаче на вход сети.

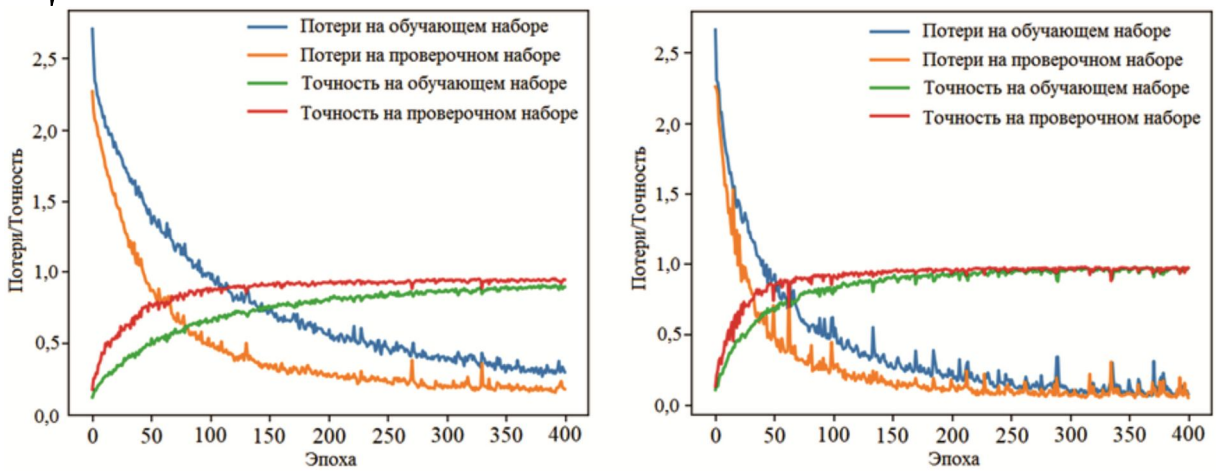
Для оценки влияния гиперпараметров трансформера на получаемые результаты были обучены и протестированы еще два его варианта, кратно отличающиеся количеством настраиваемых параметров от описанного выше.

В первом случае размерность фрагментов была увеличена с 32×32 до 64×64, то есть длина обрабатываемой последовательности уменьшилась с 16-ти до четырех, размерность слоев позиционного кодировщика снижена со 128-ти до 64-х, количество блоков трансформера уменьшено с восьми до четырех. Количество «голов» внимания не изменилось и осталось равным четырем, однако размерность векторов запроса, ключа и значения снижена с 96-ти до 64-х. Общее количество настраиваемых параметров этого варианта трансформера составило 614’090, то есть уменьшилось практически в четыре раза. Ход процесса обучения на наборе с $q = 20$ дБ иллюстрируют зависимости функции потерь и точности классификации на рисунке 6,*а*. Точность на тестовом наборе составила 91,79 %.

Во втором случае размерности фрагментов и слоев позиционного кодировщика не изменялись и составили 32×32 и 128 соответственно. Количество блоков трансформера увеличено с восьми до 16-ти. Количество «голов» внимания увеличено с четырех до восьми, размерность векторов запроса, ключа и значения увеличена с 96-ти до 128-ти. Общее количество настраиваемых параметров этого варианта сети составило 9’372’810, то есть увеличилось практически в четыре раза. Ход процесса обучения на наборе с $q = 20$ дБ иллюстрируют зависимости на рисунке 6,*б*. Точность на тестовом наборе составила 95,03 %.

Заключение

Таким образом, сравнение результатов классификации радиолокационных изображений набора MSTAR классической сверточной нейронной сетью VGG-типа и сетью с ViT-архитектурой продемонстрировало более низкое качество работы последней во всех случаях. При этом как четырехкратное увеличение, так и четырехкратное уменьшение количества настраиваемых параметров (обучаемых весов) трансформера не привело к



а) *б)*

Рисунок 6 – Ход процесса обучения сетей-трансформеров с уменьшенным до 614’090 (а) и увеличенным до 9’372’810 (б) количеством настраиваемых параметро

увеличению точности классификации по сравнению с исходным вариантом, сопоставимым по количеству весов со сверточной сетью. Сверточная сеть также продемонстрировала несколько более высокую скорость обучения.

Полученный результат, разумеется, не свидетельствует о более низком качестве работы трансформеров в задаче классификации радиолокационных изображений вообще. Но, по всей видимости, раскрытие преимуществ ViT-архитектуры по сравнению со сверточной сетью, как и в случае оптических изображений, возможно при использовании датасетов с существенно большим количеством и разнообразием примеров по сравнению с хорошо сбалансированным, однородным и относительно небольшим базовым набором MSTAR.

Литература

1. Zhu X., Montazeri S., Ali M., Hua Yu., Wang Yu., Mou L., Shi Yi., Xu F., Bamler R. Deep Learning Meets SAR. arXiv:2006.10027.

2. Chen S., Wang H., Xu F., Jin Y.-Q. Target Classification Using the Deep Convolutional Networks for SAR Images. IEEE Transaction Geoscience and Remote Sensing, 2016, vol. 54, no. 8, pp. 4806-4817. doi: 10.1109/TGRS.2016.2551720.

3. Anas H., Majdoulayne H., Chaimae A., Nabil S.M. Deep Learning for SAR Image Classification. 2020. doi: 10.1007/978-3-030-29516-5_67.

4. Chen S., Wang H. SAR Target Recognition Based on Deep Learning. 2014 International Conference on Data Science and Advanced Analytics (DSAA), 2014, pp. 541-547. doi: 10.1109/DSAA.2014.7058124.

5. Coman C., Thaens R. A Deep Learning SAR Target Classification Experiment on MSTAR Dataset. 2018 19th International Radar Symposium (IRS), 2018, pp. 1-6. doi: 10.23919/IRS.2018.
8448048.

6. Furukawa H. Deep Learning for End-to-End Automatic Target Recognition from Synthetic Aperture Radar Imagery. arXiv:1801.08558.

7. Profeta A., Rodriguez A., Clouse H.S. Convolutional Neural Networks for Synthetic Aperture Radar Classification. Proc. SPIE 9843, Algorithms for Synthetic Aperture Radar Imagery XXIII, 98430M (14 May 2016). https://doi.org/10.1117/12.2225934.

8. Amrani M., Bey A., Amamra A. New SAR Target Recognition Based on YOLO and Very Deep Multi-Canonical Correlation Analysis. arXiv:2110.15383 https://doi.org/10.48550/arXiv.2110.15383.

9. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need. arXiv:1706.03762.

10. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

11. Srivastava S., Sharma G. OmniVec: Learning robust representations with cross modal sharing. arXiv:2311.05709.

12. Li K., Zhang M., Xu M., Tang R., Wang L., Wang H. Ship Detection in SAR Images Based on Feature Enhancement Swin Transformer and Adjacent Feature Fusion. Remote Sensing 2022, 14, 3186. https://doi.org/10.3390/rs14133186.

13. Wickramasinghe S., Parikh D., Zhang B., Kannan R., Prasanna V., Busart C. VTR: An Optimized Vision Transformer for SAR ATR Acceleration on FPGA. arXiv:2404.04527.

14. Fein-Ashley J., Ye T., Kannan R., Prasanna V., Busart C. Benchmarking Deep Learning Classifiers for SAR Automatic Target Recognition. arXiv:2312.06940.

15. Купряшкин И.Ф. Сравнительные результаты точности классификации радиолокационных изображений объектов набора MSTAR сверточными нейронными сетями с различными архитектурами. Журнал радиоэлектроники, 2021, №11. https://doi.org/10.30898/1684-1719.2021.11.14.

16. https://www.sdms.afrl.af.mil/index.php?collection=mstar.

17. Шолле Ф. Глубокое обучение на Python. СПб:Питер, 2018. 400 с.

18. Simonyan K., Zisserman A. Very Deep Convolutional Networks For Large-Scale Image Recognition. arXiv:1409.1556. https://arxiv.org/abs/1409.1556.

19. Hendrycks D., Gimpel K. Gaussian Error Linear Units (GELUs). arXiv:1606.08415. https://doi.org/10.48550/arXiv.1606.08415.

20. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2018. 480 с.